心理学报 2018, Vol. 50, No.1, 130−140
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2018.00130

不同认知结构被试的测验设计模式*

彭亚风[1] 罗照盛[1] 李喻骏[2] 高椿雷[1]

([1]江西师范大学心理学院, 南昌 330022) ([2]华南师范大学心理应用研究中心/心理学院, 广州 510631)

摘 要 正如不同的病症需要使用不同的医疗技术方法来诊断一样, 不同的认知结构也需要设计对应的测验模式来进行诊断, 从而保证测验具有高质量的诊断评估效果。但传统测验形式未考虑不同认知结构的针对性诊断测验需求, 导致“千人一卷”在测验效率上有所不足; 认知诊断计算机化自适应测验虽可针对不同认知结构的被试施测不同的项目, 然而支持自适应过程的题库却没有针对不同认知结构被试设计对应的项目, 导致题库使用效率较低。要解决上述问题的关键在于, 探索如何针对不同认知结构设计相对应的测验模式。本研究采用 Monte Carlo 模拟, 对六种属性层级关系下, 不同认知结构的测验设计模式进行探讨。实验结果表明(1)同一属性层级关系下, 不同认知结构的最佳测验设计模式不同; (2)依据不同认知结构的最佳测验设计模式构建的题库具有更高的使用效率。测验编制者可以根据实验结果针对不同认知结构优化对应的测验设计模式, 并用于指导题库建设。

关键词 认知结构; 测验模式设计; 题库建设

分类号 B841

1 前言

认知诊断评价(Cognitive Diagnose Assessment, CDA, Leighton & Gierl, 2007)以认知诊断模型(Cognitive Diagnosis Model, CDM)为基础, 是对被试认知结构或认知过程的诊断评估。同其他评价方法相比, CDA 能实现对个体认知优势与劣势的诊断, 从而为教师对学生进行补救教学、开展因材施教提供指导。与此同时, 提供诊断信息, 已经成为现代教育发展的重要需求。近年来兴起的“互联网+”智慧测评, 强调测验需要向学生、家长以及教师提供诊断信息。在这个趋势下, CDA 已经成为心理与教育测量学界最重要的研究热点之一(Chiu, Douglas, & Li, 2009; de la Torre, 2008; De la Torre & Douglas, 2004; DeCarlo, 2010; Liu, Xu, & Ying, 2012; 郭磊, 苑春永, 边玉芳, 2013; 罗照盛, 李喻骏, 喻晓锋, 高椿雷, 彭亚风, 2015; 罗照盛, 喻晓锋等, 2015; 涂冬波, 蔡艳, 戴海琦, 2013; 喻晓锋等, 2015)。

正像医生需要用一些特定的医疗技术方法来诊断患者病症一样, CDA 也需要相应的工具才能探查被试不可直接观察的认知结构, 以实现其诊断功能。这个工具就是认知诊断测验(如无特别说明, 本文中的测验均指认知诊断测验)。那么如何设计一个合理的诊断测验? 一般来说, 诊断测验编制的大致流程为: 首先定义属性及其层级关系; 然后设计 Q 矩阵(表征了项目和属性间关系); 最后命题专家依据所设计的 Q 矩阵编制项目组成测验进行测试。要精确诊断不同种类的认知结构就需要使用为其“量身定制”的测验。目前, 关于诊断测验设计的研究可以分为以下两大类: 传统测验形式的设计模式研究和计算机化自适应测验(computerized adaptive test, CAT)的设计模式研究。

传统测验形式是用一套结构固定的试题去诊断具有不同认知结构的被试群体。为了实现对被试的高效诊断, 研究者就如何设计这套试题的结构进行了很多有益的探讨(Liu, Huggins-Manley, & Bradshaw, 2017; Madison & Bradshaw, 2015; 丁树良, 汪文义, 杨淑群, 2011; 丁树良, 杨淑群, 汪文义,

收稿日期: 2017-01-22

* 国家自然科学基金(31660279)、江西省社会科学规划(16JY36)、江西省研究生创新专项基金(YC2015-B025)资助。

通信作者: 罗照盛, E-mail: luozs@126.com

2010; 彭亚风, 罗照盛, 喻晓锋, 高椿雷, 李喻骏, 2016)。研究结果均指出, 在测验结构 Q 矩阵里包含 R*(由于可达矩阵是特定概念, 为不引起混淆, 本文将 Q 矩阵中包含与可达矩阵元素结构相同的矩阵子集称为类 R 阵, 记为 R*)可以提高对被试的分类准确性。进一步, 彭亚风等人(2016)针对不同属性个数及其层级关系, 提出了进行诊断评价时 Q 矩阵优化设计的一些建议。这类研究从被试群体的角度提出了传统测验形式的结构优化设计, 但是未考虑不同认知结构的针对性诊断需求, 存在“千人一卷”的相对单一性, 无法做到“因人施测”, 因而不可避免地在测验效率上有所不足。

相比之下, 认知诊断计算机化自适应测验(cognitive diagnosis computerized adaptive test, CD-CAT)有着传统测验形式所不具备的优势, 即能够根据被试认知结构的不同测试不同的项目。这种测验形式虽然能保证被试所做的项目是当前题库中最优的, 但是用来支持自适应测试过程的题库在设计时并没有考虑针对不同认知结构命制针对性的项目, 这就从根本上限制了被试与项目之间的契合程度。更进一步, 这就可能导致题库利用率出现问题, 例如, 项目的过度曝光、曝光不足以及曝光不均匀等问题。这些问题会影响测验安全, 导致项目开发与维护的成本增加(Wang, Chang, & Huebner, 2011; 毛秀珍, 辛涛, 2013; 唐小娟, 丁树良, 俞宗火, 2012)。

探索如何针对不同认知结构设计相对应的测验模式, 这是尝试解决题库建设过程中一个重要的先导问题。在解决了这一问题, 明确认知结构和项目之间关系的前提下, 才能从根本上保证“因人施测”以及优化题库设计。

综上所述, 本研究拟考察不同认知结构的测验设计模式, 以期构建出不同认知结构的最佳测验设计模式, 为题库建设提供切实可行的建议, 进而帮助提高诊断效率的同时降低题库建设成本。本文包含两个模拟实验：实验 1 探讨了不同认知结构的最佳测验设计模式; 实验 2 考察了基于不同认知结构的最佳测验设计模式在 CD-CAT 题库构建中的应用。

2　研究方法

正如前文所述, 为不同认知结构被试设计对应的测验模式是为了高效精准地诊断被试, 这与 CAT 的测验目的相吻合。而要实现这一目的需构建优质题库。为此, Reckase (2003, 2007, 2010)借助 CAT 的测验方式“反过来”探索优质题库的形态, 并提出了 CAT 中题库的优化设计方法——*p*-优化方法。Reckase 提出了最佳题库的概念, 即对每一种选题策略, 都能在题库中找到符合该策略特定范围 *p* 内的项目, 并且将这种设计称之为“*p*-优化” (*p*-optimal)。在 *p*-优化的思路下, 若采用最大 Fisher 信息量选题, 则题库中只要有能够提供达到最大信息量 *p*%的项目便可接受。据此, Reckase 提出了基于 Rash 模型的题库优化设计方法。其大致步骤为：首先, 依据 Rash 模型 Fisher 信息量的计算公式可以算出最大 Fisher 信息量 *p*%所对应的难度区间(bin), 并以此区间长度为单位将难度量表划分成多个区间; 其次, 随机抽取被试施测 CAT, 记录每个被试在每个区间上所需的项目数量; 再次, 依据最大题量原则, 对被试施测后在各个区间上的项目数量进行融合, 随着被试数量的增多, 各个区间内项目数量趋于稳定; 最后, 汇总所有区间上的项目数量及其测量学信息, 形成题库优化设计蓝图。

受 *p*-优化方法的启发, 在其基础上, 根据 CDA 的特点, 提出针对不同认知结构的最佳测验设计模式构建方法。IRT 和 CDA 存在着两点不同：第一, 项目的测量学信息不同。IRT 下表现为项目参数(如难度, 区分度), 而 CDA 的项目测量学信息不仅包括项目参数, 还包括项目所考察的属性组合。第二, 被试的测量学信息不同。IRT 中被试能力水平是连续数值, 取值范围为[−∞, +∞], 通常假定服从正态分布, 而 CDA 中被试的认知结构是离散的, 且当属性层级关系及其个数确定时, 所有可能的认知结构就是固定的, 同时典型项目考核模式也就确定了。基于 CDA 的以上特点考虑, 当测验所考察属性已确定的情况下, 将构建不同认知结构的最佳测验设计模式的具体步骤设定如下：

(1)划分区间。根据给定的属性及其层级关系, 计算典型项目考核模式, 将每种模式记为一个区间。

(2)模拟 CD-CAT 并记录各个区间内的项目数量。针对某一种认知结构的被试群体, 从中随机抽取被试进入 CD-CAT, 并记录被试所做项目。当施测完一名被试后, 根据项目所属的区间, 计算该被试在每个区间里所做项目的数量。(CD-CAT 的题库中包含所有典型项目考核模式, 且每种模式的项目数量足够大, 项目参数的分布区间足够广。)

(3)融合。依据最大题量原则, 对被试施测后在各个区间上的项目数量进行融合。例如, 在区间 A

上，被试 1 测了 5 个项目，被试 2 为 3 个项目，融合后以 5 个项目作为区间 A 的期望项目数量。

(4)重复第 2 步和第 3 步，直到该认知结构中所有被试施测完毕，汇总各个区间内的项目数量，即可得到该种认知结构下的最佳测验设计模式。

下面通过一个例子进行简单的说明。

假定属性个数为 5，属性层级关系为独立型(Independent) (Tatsuoka, 1995)。被试 *i* 和被试 *n* 的认知结构真值均为[1 1 0 0 0]。采用后验加权 K-L 信息(posterior-weighted Kullback-Leibler, PWKL) (Cheng, 2009)选题，直到被试 *i* 和被试 *n* 在某种认知结构的最大后验概率不低于 0.95，则终止测验。记录两名被试在整个测验过程中的选题、作答及参数估计情况，结果如表 1 和表 2 所示。

表 1 被试 *i* 基于 PWKL 所选出的项目、作答、认知结构估计值及其后验概率

项目顺序	项目属性向量	作答	认知结构估计值	后验概率
1	[0 1 1 0 0]	0	[0 0 0 0 0]	0.0408
2	[0 0 0 0 1]	0	[0 0 0 0 0]	0.0772
3	[1 0 0 0 0]	1	[1 0 0 0 0]	0.1321
4	[1 0 0 1 0]	0	[1 0 0 0 0]	0.2219
5	[0 1 0 0 0]	1	[1 1 0 0 0]	0.5687
6	[1 1 0 0 0]	1	[1 1 0 0 0]	0.8187
7	[0 0 1 0 0]	0	[1 1 0 0 0]	0.8725
8	[0 1 0 1 0]	0	[1 1 0 0 0]	0.9236
9	[0 1 0 0 1]	0	[1 1 0 0 0]	0.9719

表 2 被试 *n* 基于 PWKL 所选出的项目、作答、认知结构估计值及其后验概率

项目顺序	项目属性向量	作答	认知结构估计值	后验概率
1	[0 1 1 1 0]	0	[0 0 0 0 0]	0.0354
2	[0 0 0 0 1]	0	[0 0 0 0 0]	0.0661
3	[0 0 0 1 0]	1	[0 0 0 1 0]	0.1356
4	[1 0 0 1 0]	0	[0 0 0 1 0]	0.2339
5	[0 0 1 0 0]	0	[0 0 0 1 0]	0.3618
6	[0 1 0 0 0]	1	[0 1 0 1 0]	0.6818
7	[0 1 0 1 0]	0	[0 1 0 0 0]	0.2269
8	[0 0 0 1 0]	0	[0 1 0 0 0]	0.3832
9	[1 1 0 0 0]	1	[1 1 0 0 0]	0.786
10	[1 1 0 0 0]	1	[1 1 0 0 0]	0.8715
11	[1 0 0 0 1]	1	[1 1 0 0 1]	0.5342
12	[1 1 0 0 1]	0	[1 1 0 0 0]	0.8578
13	[1 1 0 0 1]	0	[1 1 0 0 0]	0.9272
14	[1 0 1 0 0]	0	[1 1 0 0 0]	0.979

表 1 和表 2 分别为被试 *i* 和被试 *n* 在 CD-CAT 过程中所选择项目的属性向量、作答情况、做完每个项目后认知结构的估计值及其后验概率。按照最大题量的原则对每个区间内的项目个数进行融合，得到表 3。

表 3 各个区间内抽取的项目个数(融合后)

区间	项目个数
[1 0 0 0 0]	1
[0 1 0 0 0]	1
[0 0 1 0 0]	1
[0 0 0 1 0]	2
[0 0 0 0 1]	1
[1 1 0 0 0]	2
[1 0 1 0 0]	1
[1 0 0 1 0]	1
[1 0 0 0 1]	1
[0 1 1 0 0]	1
[0 1 0 1 0]	1
[0 1 0 0 1]	1
[1 1 0 0 1]	2
[0 1 1 1 0]	1

同时，通过表 1 和表 2 可以发现单个被试的整个测验过程可以分为两个阶段：试验性探查阶段(记为 0 阶段)和精确估计阶段(记为 1 阶段)。其中，1 阶段指的是被试认知结构的估计值与真值一致之后的阶段，在此阶段所做项目不会改变认知结构的估计值，只会一直增加该估计值的后验概率直至达到终止规则，例如表 1 中的 5~9 题；0 阶段就是 1 阶段之前的阶段，例如表 1 中的 1~4 题，表 2 中的 1~11 题。需要注意的是，受随机因素的影响，0 阶段时被试认知结构的估计会存在波动，例如表 2 中被试 *n* 在完成 9~10 题后，其认知结构均估计正确，但第 11 题之后又估计错误。

那么，不同认知结构的被试在 0 阶段和 1 阶段抽取的项目类型是否存在某种规律？探讨此问题可为不同认知结构的最佳测验设计模式提供更加明确的设计方向，进一步节省测验编制的成本。

3 实验 1：不同认知结构的最佳测验设计模式

本研究考察在不同属性个数、不同属性层级关系的情况下，不同认知结构的最佳测验设计模式。

3.1 方法

3.1.1 属性个数及其层级关系的类型

本研究考察的属性个数有两种水平：*K*=5 个、

K=6。属性层级关系：6 种类型，分别为：直线型(Linear)、收敛型(Convergent)、发散型(Divergent)、无结构型(Unstructured)、独立型(Independent)、混合型(Mixture)。其中，混合型是一种多种属性层级关系并存的关系类型，是为了仿真实际测验情境中可能存在较为复杂的属性层级关系模式。(所有属性层级关系示意图见网络版附录 1，附录 2)。

3.1.2　被试与题库设计

(1)被试设计

由于当属性的个数及其层级关系确定以后，所有认知结构的类型便以确定。因此，为了探索出每种认知结构的最佳测验设计模式，固定每种认知结构的被试人数均为 100 人。即被试认知结构分布，服从均匀分布。

(2)题库设计

属性的个数及其层级关系决定了典型项目考核模式的种类。因此，设定每种典型项目考核模式均重复出现 40 次。项目参数：s 和 g 服从均分分布 U (0.05, 0.25)。

3.1.3　采用的认知诊断模型

本研究采用的认知诊断模型为 DINA 模型。DINA 模型在拥有简洁项目参数的同时，分类准确性较高(De la Torre & Douglas, 2004)。DINA 模型的公式如下：

$$P\left(Y_{ij}=1\middle|\alpha_i\right)=\left(1-s_j\right)^{\eta_{ij}}g_j^{1-\eta_{ij}} \tag{1}$$

其中，

$$\eta_{ij}=\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}} \tag{2}$$

η_{ij} 表示的是被试 i 是否掌握了项目 j 所考核的所有属性；q_{ij} 表示的是项目 j 是否考察了属性 k；s_j：被试掌握了项目 j 所考核的所有属性，但答错的概率；g_j：被试未全部掌握项目 j 所考核的所有属性，但是答对的概率。

3.1.4　采用的选题策略

PWKL 为判准率较高且使用较为广泛的一种选题策略，其公式如下：

$$PWKL_j\left(\hat{\alpha}_i^t\right)=\sum_{c=1}^{2^K}\left\{\sum_{y=0}^{1}\left[\log\left(\frac{P\left(Y_{ij}=y\middle|\hat{\alpha}_i^t\right)}{P\left(Y_{ij}=y\middle|\alpha_c\right)}\right)P\left(Y_{ij}=y\middle|\hat{\alpha}_i^t\right)\right]\pi_{i,t}\left(\alpha_c\right)\right\} \tag{3}$$

其中，$\hat{\alpha}_i^t$ 为被试 i 作答完 t 个项目后认知结构的估计值，$P\left(Y_{ij}=y\middle|\hat{\alpha}_i^t\right)$ 是认知结构为 $\hat{\alpha}_i^t$ 的被试在项目 j 上作答反应是 y 的概率；α_c 为任意一种认知结构(c=1, 2, 3,⋯, 2^K)，$P\left(Y_{ij}=y\middle|\alpha_c\right)$ 是认知结构为 α_c 的被试在项目 j 上作答反应是 y 的概率；$\pi_{i,t}\left(\alpha_c\right)$ 是认知结构为 α_c 后验概率。PWKL 的选题标准为在当前认知结构估计值 $\hat{\alpha}_i^t$ 下，从剩余题库中选择具有最大 PWKL 值的项目给被试作答。

3.1.5　终止规则

采用变长终止规则，设定的标准为当被试属于某种认知结构的最大后验概率不低于 0.95。使用变长的终止规则来探索每种认知结构的最佳测验设计模式的原因在于，一方面可以设定每个被试的测量精度相同，可以有更高的估计精度(Babcock & Weiss, 2009)；另一方面更能体现出自适应的特点和优势(郭磊，郑蝉金，边玉芳, 2015)。

3.1.6　CD-CAT 模拟及认知结构估计

用 Monte Carlo 方法进行模拟。第 2 部分已经详细地介绍了实验的具体过程，这里主要介绍第 2 部分步骤 2 中被试 CD-CAT 过程的模拟方法，具体步骤如下：

(1)选题。使用 PWKL 选题，随机产生被试的认知结构初值，然后基于认知结构初值通过 PWKL 选择第一个项目。

(2)模拟作答。运用 DINA 模型的项目反应函数计算被试在所选项目上的正确作答概率 p。然后生成一个随机数 r，若 p>r，则被试在该项目上的作答记为 1，否则为 0 分。

(3)估计认知结构。根据被试在已作答项目上的反应通过最大后验概率方法(Maximum A Posterior, MAP)估计被试的认知结构及其后验概率。

(4)再选题。再根据选题策略从剩余题库中选出与被试当前认知结构估计值最匹配的项目给被试作答。

(5)重复步骤 2 至步骤 4 直至被试属于某种认知结构的最大后验概率不低于 0.95。

在每个被试 CD-CAT 模拟过程中，记录其所抽取的项目、每做完一个项目后认知结构的估计值及其后验概率。

实验重复次数为 30 次。

3.1.7　评价指标

模式判准率(Pattern Match Ration, PMR)用于考察被试认知结构的仿真性，它指被试认知结构判对的人数占总人数的百分比，PMR 越大，表明分类准确性越高。计算公式如下：

$$PMR=\frac{\sum_{i=1}^{N}N_{i_correct}}{N} \tag{4}$$

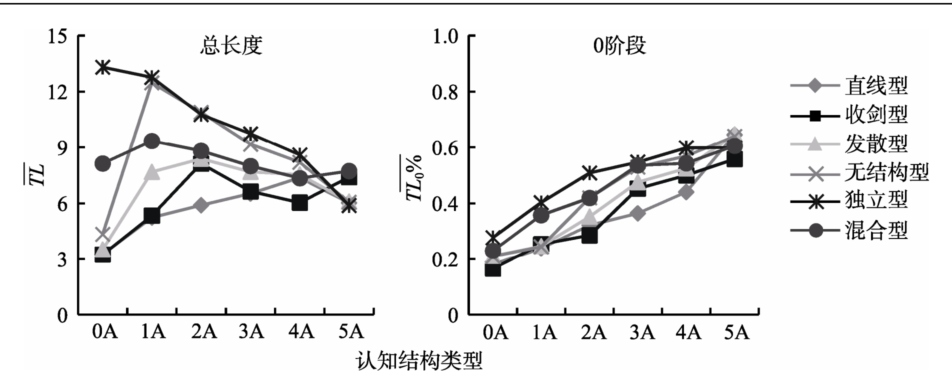


图 1　六种属性层级关系下不同认知结构类型的平均测验总长度以及 0、1 阶段测验长度占总长度百分比(K=5, 30 次平均结果)

其中, N 为被试总人数。若被试 i 的认知结构真值与其参加测验后估计得到的认知结构估计值相等，则 $N_{i_correct}=1$，反之则 $N_{i_correct}=0$

3.2　实验结果

六种属性层级关系下使用 PWKL 选题的 PMR 均在0.96之上，具有较高判准率。图 1 为 K=5 时，六种属性层级关系下每种认知结构的平均测验长度(记为 $\overline{TL}$)以及 0、1 阶段测验长度占总长度的百分比(分别记为 $\overline{TL_0}\%$ 和 $\overline{TL_1}\%$) (K=6 时的结果呈现相同趋势，请见网络版附录 3)，横坐标为认知结构的种类，纵坐标为平均测验长度。因为属性层级关系越松散，其对应的认知结构的种类就越多，故为了便于结果的清晰呈现，依据认知结构中掌握的属性个数将每种属性层级关系下所有的认知结构分为 6 类：掌握了 0、1、2、3、4、5 个属性的认知结构类型，分别记为 0A, 1A, 2A, 3A, 4A, 5A (下同)。

总体而言，属性层级关系越紧密，$\overline{TL}$ 越小；同一属性层级关系下不同认知结构类型的 $\overline{TL}$ 各不相同。六种属性层级关系下, 0 阶段时的平均测验长度与认知结构中掌握的属性个数成正比($\overline{TL_0}\%$ 逐渐增大)，1 阶段与之相反($\overline{TL_1}\%$ 逐渐减小)。即 0A 至 5A，$\overline{TL_0}\%$ 大致范围分别为：20%~30%、20%~40%、30%~50%、40%~60%、60%左右，$\overline{TL_1}\%$ 大致范围分别为 1− $\overline{TL_0}\%$ 。

以直线型为例，对 1 阶段时认知结构为[0 0 0 0 0]的所有被试在区间[1 0 0 0 0]上抽取的项目数量进行频次分析，结果见表 4。通过表 4 发现，在区间[1 0 0 0 0]上抽取了 2 个项目的有 86 人次，抽取了 3 个项目的有 6 人次，项目的有 9 人次，抽取了 6 个项目的仅有 1 人次，此时最大题量为 6。由此可见，第 2 部分步骤 3 并不适合使用最大题量的原则进行融合。因为若按照最大题量的融合原则，则[0 0 0 0 0]的认知结构在区间[1 0 0 0 0]上需要 6 个项目，而

表 4　认知结构为[0 0 0 0 0]的所有被试在区间[1 0 0 0 0]上抽取项目数量的频次分布

项目数量	人数
2	84
3	6
4	9
5	0
6	1

实际上，绝大多数被试都是抽取了 3 个以下的项目，这会造成这个区间内的项目数量虚高，增加命题成本。因此，本研究采用了另外两种方法来进行融合。方法 1：区间内项目数量分布的平均数加 1 个标准差(记为 $M+SD$)；方法 2：区间内项目数量分布的第 90 百分位数(记为 p_{90})。

图 2 为 K=5 时，使用 p_{90} 得到的六种属性层级关系下，所有认知结构类型在 0、1 阶段选出来的项目类型及其个数($M+SD$ 的结果与 p_{90} 基本一致，见网络版附录 4; K=6 时呈现相同趋势，对应结果请见网络版附录 5 和附录 6)。为了结果的清晰呈现，对实验结果进行如下处理：首先，对每个区间按照其考察的属性个数进行分类，分为：考察 1、2、3、4、5 个属性的项目类型，分别记为 1IA, 2IA, 3IA, 4IA, 5IA (下同)；然后，针对每种认知结构，将各个项目类型所包含区间里的项目数量分别累加；最后，求取各认知结构类型下，上一步所得累加值的平均数。结果如图 2 所示。

从图 2 中的每一行可以看出，同一属性层级关系内不同认知结构类型抽取的项目类型及其个数均不同，即不同认知结构有不同的最佳测验设计模式。认知结构类型中掌握的属性个数与抽取的项目类型中考察的属性个数呈正比。0 阶段和 1 阶段最佳测验设计模式的相同之处在于，项目的抽取围绕目标属性(当前认知结构中掌握的属性)展开，且随

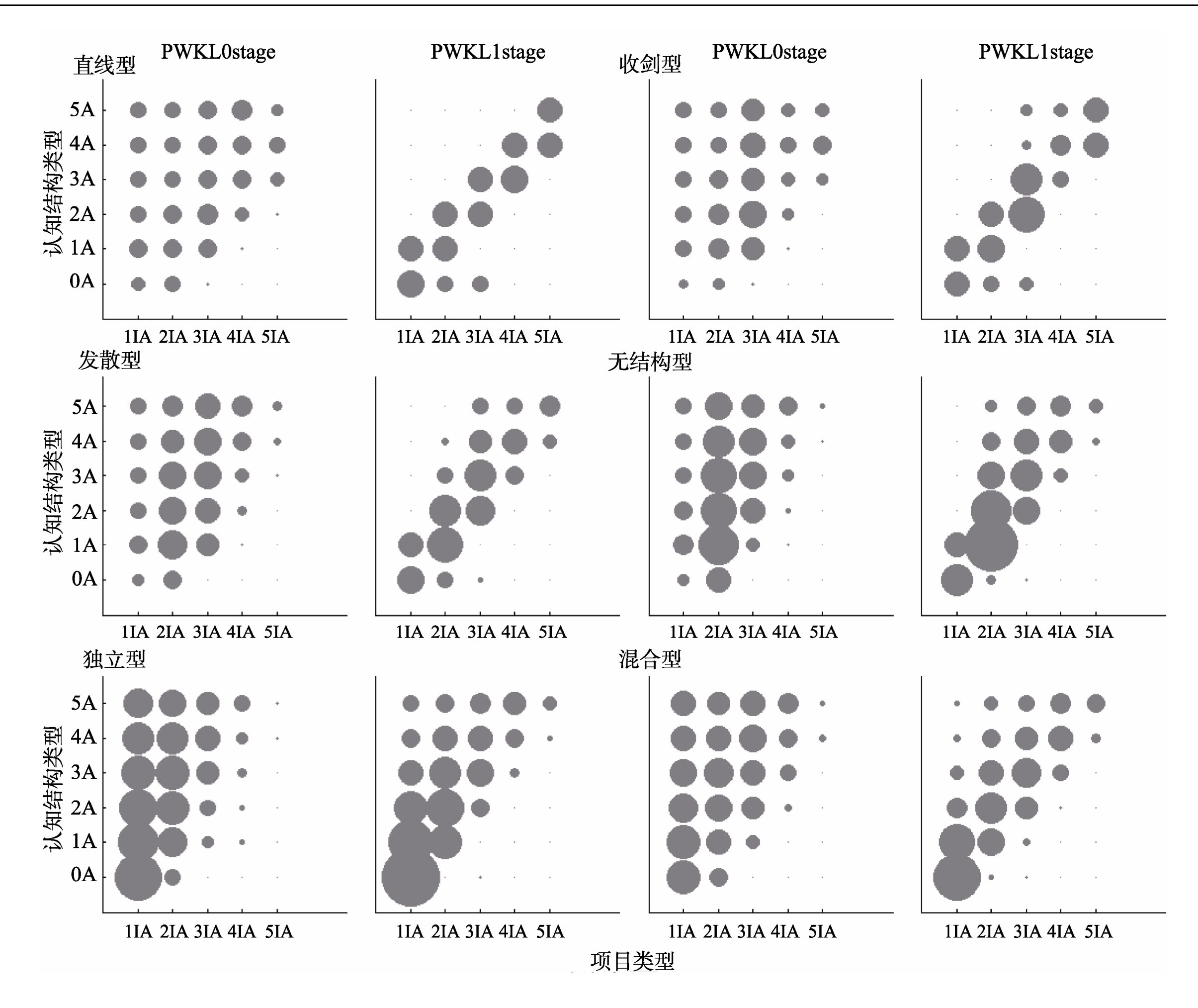


图 2　六种属性层级关系下所有认知结构类型在 0、1 阶段下选出的项目类型及其个数(K=5, p_{90}, 30 次平均结果)。图中圆圈面积的大小与项目个数成正比：项目被抽取的个数越多, 图中圆圈对应的面积就越大。最小面积代表被抽取的项目平均个数为 0.2, 最大面积代表被抽取的项目平均个数为 12.9。

着认知结构类型中掌握属性个数的增加, 抽取的项目类型种类增多；不同之处在于, 0 阶段时, 不同认知结构类型抽取的项目类型大部分比较一致, 1 阶段更多抽取的是考察属性个数与认知结构类型中掌握属性个数较为接近的项目类型(图中表现为明显的对角线集中趋势)。

具体来说, 0 阶段时, 随着认知结构类型中掌握属性个数的增加, 3IA、4IA 中的项目个数逐渐增加：0A 时抽取 1IA 中每种项目考核模式 1~2 个, 2IA 中每种项目考核模式 1~2 个；1A 时抽取 1IA、2IA 中考察了目标属性的项目 2 个, 1IA 中考察非目标属性的项目各 1 个；2A 在 1A 基础上增加 2IA、3IA 中考察目标属性的项目各 1 个；3A 在 1A 基础上增加 2IA 中未考察目标属性的项目 2 个以及 3IA 中考察目标属性的项目各 2 个；4A、5A 抽取的项目类型与 3A 大致相同。1 阶段时, 0A 时抽取 1IA 中每种项目考核模式 2~3 个；1A~5A 的最佳测验模式为在 0 阶段对应模式基础上, 逐渐减少了 1IA、2IA 的项目个数, 逐渐增加了 3IA、4IA、5IA 中考察目标属性的项目个数 1~2 个。

Flaugher (2000)指出要实现 CAT 的优势, 题库中必须包含针对不同能力水平的高质量题目。同理, CD-CAT 可以为每种认知结构提供最匹配测验的前提是, 题库中应该包含针对每种认知结构类型所需的所有项目类型及其个数。因此, 使用最大题量原则将所有认知结构的最佳测验设计模式在每个区间内的项目数量进行融合, 得到了该种属性个数及其层级关系下的题库建设蓝图, 结果见表 5。

表 5 为六种属性层级关系下题库中需要的项目类型及其个数。从表 5 可以看出六种属性层级关系下, 题库容量与层级关系的紧密程度成反比。因为属性层级关系越紧密, 其对应的典型项目考核模式的种类就越少, 从而导致每种项目类型里的项目数量也越小。

表 5 不同属性个数时六种属性层级关系下题库中各个项目类型的数量分布

属性层级关系	属性个数	项目类型						题库容量
		1IA	2IA	3IA	4IA	5IA	6IA	
直线型	K=5	5	4	4	4	4		21
	K=6	4	5	5	5	5	4	28
收敛型	K=5	5	5	9	5	4		28
	K=6	4	5	10	4	5	4	32
发散型	K=5	5	10	11	8	2		36
	K=6	6	10	17	14	6	2	55
无结构型	K=5	5	18	16	10	2		51
	K=6	7	22	27	21	5	1	83
独立型	K=5	22	28	19	8	1		78
	K=6	30	47	38	21	7	0	143
混合型	K=5	14	15	12	8	2		51
	K=6	15	13	12	12	8	2	62

具体来说(以 K=5 为例), 从题库容量和属性层级关系的对应关系上来看, 每种属性层级关系下的典型项目考核模式种类是决定题库容量的重要指标。直线型、收敛型和发散型下所需的题库容量是对应的典型项目考核模式种类的 4~5 倍, 无结构和混合型时为 3~4 倍, 独立型为 2~3 倍。例如, 独立型情况下典型项目考核模式有 31 种, 则此时的题库容量为 62~93 之间较为合适。

进一步地, 从题库和项目类型的关系上看, 不同的项目类型有着不同的项目数量, 影响着题库的大小。每种项目类型的项目数量与该项目类型所包含典型项目考核模式的种类有关, 且因属性层级关系的不同而不同: 直线型和收敛型情况下, 1IA~5IA 中所包含的每种典型项目考核模式均 5 个左右, 例如, 直线型情况下 1IA 中仅包含 1 种典型项目考核模式([1 0 0 0 0]), 则题库中应该包含 5 个该种考核模式的项目, 最终 1IA 的项目数量为 5, 其余情况以此类推; 发散型下, 1IA 至 5IA 中所包含的每种典型项目考核模式的项目个数分别为 5 个, 5 个, 4 个, 4 个, 2 个; 当属性层级关系为无结构时, 分别为: 5 个, 5 个, 3 个, 3 个, 2 个; 独立型情况下, 对应的项目个数分别为: 5 个, 3 个, 2 个, 2 个, 1 个。混合型时 1IA 至 5IA 中所包含的每种典型项目考核模式的项目个数验证了上述结果, 例如 1IA 中每种典型项目考核模式为 5 个; 2IA 中属于独立型关系的属性(A1 和 A4)组成的典型项目考核模式([1 0 0 1 0])的项目个数和属于直线型关系的属性(A1 和 A2)组合([1 1 0 0 0])的项目个数分别为 3 个和 5 个; 3IA 中属于独立型关系的属性(A1、A4 和 A5)组合([1 0 0 1 1])的项目个数为 2 个, 属于收敛型关系的属性(A1、A3 和 A4)组合([1 0 1 1 0])的项目个数为 4 个。

实验 1 还在 CD-CAT 中使用了香农熵(Shannon Entropy, SHE) (Tatsuoka, 2002; Xu, Chang, & Douglas, 2003)选题策略, 实验结果呈现出相同的规律(限于篇幅未在本文中列出, 感兴趣的读者, 可与作者联系)。

4 实验 2: 基于不同认知结构的最佳测验设计模式在 CD-CAT 题库构建中的应用

目前研究者常用的两个题库模拟方法: 陈平提出的模拟题库的方法(Chen, Xin, Wang, & Chang, 2012; 陈平, 2011; 陈平, 辛涛, 2011a, 2011b)以及 Cheng 的方法(Cheng, 2009, 2010; Zheng & Chang, 2016; 毛秀珍, 辛涛, 2013)。实验 2 的主要目的是比较这两种题库与实验 1 中基于不同认知结构的最佳测验设计模式构建的题库在 CD-CAT 中的使用效率。

4.1 方法

属性个数: K=6。属性层级关系为独立型。采用的认知诊断模型为 DINA 模型, 选题策略为 PWKL。CD-CAT 模拟及认知结构估计与实验 1 一致。实验重复次数为 30 次。

4.1.1 被试与题库设计

被试总人数为 1000, 并且假设每个被试掌握每个属性的概率是 50%。题库的生成: 包含 3 个题库, 分别是: 题库 1 按照实验 1 中得到的独立型 K=6 时题库建设规律生成, 题库容量为 152, 其中 1IA-6IA 里每种项目考核模式的项目个数为: 5 个、3 个、2 个、2 个、1 个、1 个; 题库 2 按照 Cheng 的方法: 每个项目至少考查一个属性, 并且考查每个属性的概率为 0.2, 题库大小与题库 1 一致; 题库 3 按照陈平的方法生成一个 360×6 的 Q 矩阵, 其中包含三种类型的基本 Q 矩阵。题库中的项目参数 s 和 g 服从均分分布 U (0.05, 0.25)。

4.1.2 终止规则

分别采用定长与变长的终止规则: 定长下设定测验长度 TL 为 20; 变长下设定被试属于某种认知结构的最大后验概率不低于 0.95。

4.1.3 评价标准

(1)被试诊断效果评价指标: 采用重复实验下的 PMR 来评价诊断效果;

(2)题库使用均匀性指标(χ^2)：χ^2 用于评价项目观察曝光率和期望曝光率之间的差异，其计算公式如下：

$$\chi^2 = \frac{\sum_{j=1}^{J}\left(er_j - \overline{er_j}\right)}{\overline{er_j}} \tag{5}$$

其中，er_j 是第 j 个项目的曝光率，等于作答项目 j 的被试人数除以参加测验的总被试人数 N；$\overline{er_j}$ 为项目 j 的期望曝光率，等于测验长度 TL 除以题库容量。χ^2 指标越小，说明整个题库的使用越均匀。

(3)测验重叠率指标(记为 $\hat{T}$)：随机选择的两个被试之间期望重叠的项目个数与测验长度之比，其计算公式如下：

$$\hat{T} = \frac{\sum_{j=1}^{J} T_j\left(T_j - 1\right)}{TL * N * \left(N - 1\right)} \tag{6}$$

其中，T_j 是第 j 个项目的被调用次数，其余符号的定义与 χ^2 相同。

此外，还记录了最大曝光率(记为 *Max.er*)、最小曝光率(记为 *Min.er*)、曝光率大于 20%的项目数量(记为 *er* ≥ 20%)以及题库中未使用的项目数量百分比(记为 *never used%*)。

4.2　实验结果

研究结果如图 3、表 6 和表 7 所示。图 3 呈现的是 CD-CAT 中不同终止规则下不同题库对被试的诊断效果，表 6 和表 7 分别呈现的是 CD-CAT 中不同终止规则下 3 种题库的题库使用情况指标。

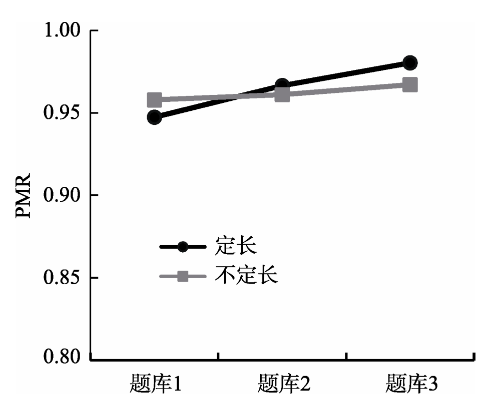


图 3　不同终止规则下 3 种题库的 PMR

表 6　定长情况下 3 种题库的使用情况

题库	χ^2	$\hat{T}$	*Max.er*	*Min.er*	*er* ≥ 20%	never used%
题库 1	45.96	0.46	0.96	0	32.23	7.19%
题库 2	58.00	0.54	0.99	0	32.36	24.3%
题库 3	150.43	0.47	0.96	0	30.00	52.06%

表 7　变长情况下 3 种题库的使用情况以及平均测验长度

题库	平均测验长度	*Max.er*	*Min.er*	*er* ≥ 20%	never used%
题库 1	15.33	0.94	0	22.43	15.11%
题库 2	14.32	0.98	0	20.73	24.93%
题库 3	12.99	0.96	0	18.1	58.61%

从图 3 可以看出，在定长与变长两种终止规则下，3 种题库的 PMR 差异不大，基本均在 0.95 之上。

由表 6 可知，题库 1 的 χ^2 指标最小为 45.96，题库 3 的 χ^2 指标最大；就重叠率而言，题库 1 的重叠率最低，题库 2 的最高；另外，三种题库的最大项目曝光率都接近 1，最小项目曝光率均为 0，曝光率大于 20%的项目均在 30 个左右。值得注意的是，题库 1 的未使用项目比例最低，仅为 7.19%。

由表 7 可知，题库 1 的平均测验长度为 15.33，略高于题库 2 的 14.32 和题库 3 的 12.99；三种题库的最大项目曝光率都接近 1，最小项目曝光率均为 0，曝光率大于 20%的项目均在 20 个左右。同样的，题库 1 的未使用项目比例最低，为 15.11%，题库 3 最高，为 58.61%。

从整体上看，题库 1 的使用效率最高，在题库使用方面的表现较其他两种题库要好。

5　讨论

5.1　不同认知结构的最佳测验设计模式

从实验结果可以看出，不同认知结构的最佳测验设计模式不相同。具体表现在不同认知结构的最佳测验长度，试验性探查阶段(0 阶段)和精确估计阶段(1 阶段)的设计模式均不相同。其中，最佳测验长度由 0、1 阶段的最佳测验设计模式所决定。0 阶段是对被试认知结构的试验性探查阶段，需要逐个排查被试在每个属性上是否掌握。因此，不同认知结构类型测验设计模式中的大部分项目类型比较一致；而 1 阶段是对被试认知结构的精确估计阶段，对不同认知结构有着更加精确的定位需求，与之对应的最佳测验设计模式也呈现出更加明显的特点：不同认知结构类型的最佳测验设计模式中的项目，其考察的属性个数与当前认知结构类型中掌握的属性个数较为接近。与此同时，0 阶段和 1 阶段的最佳测验设计模式也有共同之处，即均围绕目标属性展开。具体规律如下：

0 阶段时，掌握了 0 个属性的认知结构(记为 0A，以此类推，随着认知结构中掌握的属性个数的

增加，分别记为 1A，2A，……，*K*A)的最佳测验设计模式为：考察 1 个属性的每种项目考核模式和考察 2 个属性的项目考核模式各 1~2 个，例如直线型情况下，认知结构[0 0 0 0 0]需要考核模式为[1 0 0 0 0]和[1 1 0 0 0]的项目各 1~2 个；1A：考察 1 个和 2 个属性的项目类型中考察了认知结构已掌握属性(目标属性)的项目各 2 个，考察 1 个属性的项目类型中考察了认知结构未掌握属性(非目标属性)的项目各 1 个，例如独立型情况下，认知结构[1 0 0 0 0](第一个属性为目标属性，其余为非目标属性)需要考核模式为[1 0 0 0 0]的项目 2 个，[1 1 0 0 0]、[1 0 1 0 0]、[1 0 0 1 0]、[1 0 0 0 1]这四种考核模式中的任意 1 种 2 个或任意 2 种各 1 个，以及[0 1 0 0 0]、[0 0 1 0 0]、[0 0 0 1 0]、[0 0 0 0 1]这四种考核模式各 1 个；2A：在 1A 基础上增加考察 2、3 个属性的项目类型中考察了目标属性的项目各 1 个；3A 在 1A 基础上增加考察 2 个属性的项目类型中考察了非目标属性的项目 2 个，考察 3 个属性的项目类型中考察目标属性的项目各 2 个；4A~*K*A：在(*K*-1)A 基础上，增加考察 *K* 个属性的项目 1 个左右。

1 阶段时，随着认知结构中掌握属性个数的增加，考察 1 个和 2 个属性的项目个数逐渐减少，考察属性向量与认知结构属性向量相同以及与其相差 1~2 个属性的项目类型逐渐增多，0A 除外。0A 时抽取考察 1 个属性的项目类型中每种项目考核模式 2~3 个；1A~*K*A 的最佳测验设计模式为在 0 阶段对应模式基础上，减少了考察 1、2 个属性的项目个数，相应增加考察 3、4 至 *K* 个属性中考察目标属性的项目个数 1~2 个。

通过将测验过程划分为 0 阶段和 1 阶段可以看出，在不同的测验情景下，0 阶段的最佳测验设计模式都具有一定的共性，结合 1 阶段最佳测验设计模式所体现出的特异性，能够为题库蓝图设计提供更加明确的指导意见，进一步的节约命题成本。

5.2 基于不同认知结构的最佳测验设计模式建构题库

实验 2 的结果可以看出，基于不同认知结构的最佳测验设计模式构建出的题库，其使用效率比研究者常用的题库更高。这表明采用实验 1 得到的题库蓝图可以指导题库的建设，缓解了题库中项目浪费的情况。通过分析实验 1 中的题库蓝图，推论得到了不同属性层级关系下题库建设的一般规律：

在题库容量方面，目标领域内属性个数及其层级关系下的典型项目考核模式种类是决定题库容量的重要指标。直线型、收敛型和发散型下所需的题库容量是对应的典型项目考核模式种类的 4~5 倍，无结构时为 3~4 倍，独立型为 2~3 倍。

在题库所包含的项目类型方面，每种项目类型的项目数量与该项目类型所包含典型项目考核模式的种类有关，且因属性层级关系的不同而不同：直线型和收敛型情况下，每种项目类型中所包含的每种典型项目考核模式均 5 个左右；剩下三种属性层级关系下，随着项目类型中考察的属性个数的增加，对应所包含的每种典型项目考核模式的项目个数依次减少：考察 1~3 个属性的项目类型中每种典型项目考核模式的项目个数分别为 5、4、3 个左右，考察 4 个至 *K*–1 个属性的项目类型中每种典型项目考核模式的项目个数均为 2 个左右，以及 1 个左右考察 *K* 个属性的典型项目考核模式。

综上所述，本研究通过探讨每种认知结构的最佳测验设计模式，明确了认知结构与项目类型之间的关系，找到不同认知结构所需的针对性项目，并在此基础上推论得到题库蓝图建设的一般规律。

建设题库是一项系统工程，需要多学科专业人员(学科专家、心理与教育测量人员、计算机技术人员等)协同攻关，在科学的题库建设理论指导下有步骤地进行(漆书青，戴海琦，丁树良，2002)。本文从理论上探讨了题库建设的一般框架，提供一种科学建设题库的新方法。实践者可以依据该新方法，通过模拟事先确定题库的大致结构，再根据实际需要结合考察学科内容、测验时间等因素，进一步细化题库建设方案，这样构建出的题库既适用于诊断包含有不同认知结构类型的被试群体，又同时避免了命制实则无法助益于提升测验效率的项目，节约题库建设成本。但该题库容量相对较小，在一定程度上会增大项目过度曝光的可能性，这也是之后研究所需改进的方向。

参 考 文 献

Babcock, B., & Weiss, D. J. (2009). *Termination criteria in computerized adaptive tests: Variable-length cats are not biased*. Paper presented at the Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.

Chen, P. (2011). *Item replenishing in cognitive diagnostic computerized adaptive testing- based on DNA model* (Unpublished doctorial dissertation). Beijing Normal University.

[陈平. (2011). *认知诊断计算机自适应测验的项目增补——以 DINA 模型为例* (博士学位论文). 北京师范大学.]

Chen, P., & Xin, T. (2011a). Developing on-line calibration methods for cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica, 43*(6), 710–724.

[陈平，辛涛．(2011a)．认知诊断计算机化自适应测验中在线标定方法的开发．*心理学报, 43*(6), 710–724.]

Chen, P., & Xin, T. (2011b). Item replenishing in cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica, 43*(7), 836–850.

[陈平，辛涛．(2011b)．认知诊断计算机化自适应测验中的项目增补．*心理学报, 43*(7), 836–850.]

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika, 77*(2), 201–222.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619–632.

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement, 70*, 902–913.

Chiu, C.-Y., Douglas, J. A., & Li, X. D. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633–665.

De la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

De la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA Model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*(1), 8–26.

Ding, S. L., Wang, W. Y., & Yang, S. Q. (2011). The design of cognitive diagnostic test blueprints. *Journal of Psychological Science, 34*(2), 258−265.

[丁树良，汪文义，杨淑群. (2011). 认知诊断测验蓝图的设计．*心理科学, 34*(2), 258−265.]

Ding, S. L., Yang, S. Q., & Wang, W. Y. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *Journal of Jiangxi Normal University (Natural Science), 34*(5), 490−494.

[丁树良，杨淑群，汪文义. (2010). 可达矩阵在认知诊断测验编制中的重要作用．*江西师范大学学报(自然科学版), 34*(5), 490−494.]

Flaugher, R. (2000). Item pools. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, …D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 37–59). Mahwah, NJ: Lawrence Erlbaum Associates.

Guo, L., Yuan, C. Y., & Bian, Y. F. (2013). Discussing the development tendency of cognitive diagnosis from the perspective of new models. *Advances in Psychological Science, 21*(12), 2256–2264.

[郭磊，苑春永，边玉芳．(2013)．从新模型视角探讨认知诊断的发展趋势．*心理科学进展, 21*(12), 2256–2264.]

Guo, L., Zheng, C. J., & Bian, Y. F. (2015). Exposure control methods and termination rules in variable-length cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica, 47*(1), 129–140.

[郭磊，郑蝉金，边玉芳．(2015)．变长 CD-CAT 中的曝光控制与终止规则．*心理学报, 47*(1), 129–140.]

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge UK: Cambridge University Press.

Liu, J. C., Xu, G. J., & Ying, Z. L. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*(7), 548–564.

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement, 77*(2), 220–240.

Luo, Z. S., Li, Y. J., Yu, X. F., Gao, C. L., & Peng, Y. F. (2015). A simple cognitive diagnosis method based on Q-matrix theory. *Acta Psychologica Sinica, 47*(2), 264–272.

[罗照盛，李喻骏，喻晓锋，高椿雷，彭亚风. (2015). 一种基于 Q 矩阵理论朴素的认知诊断方法．*心理学报, 47*(2), 264–272.]

Luo, Z. S., Yu, X. F., Gao, C. L., Li, Y. J., Peng, Y. F., Wang, R., & Wang, Y. T. (2015). Item selection strategies based on attribute mastery probabilities in CD-CAT. *Acta Psychologica Sinica, 47*(5), 679–688.

[罗照盛，喻晓锋，高椿雷，李喻骏，彭亚风，王睿，王钰彤. (2015). 基于属性掌握概率的认知诊断计算机化自适应测验选题策略．*心理学报, 47*(5), 679–688.]

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75*(3), 491–511.

Mao, X. Z., & Xin, T. (2013). A comparison of item selection methods for controlling exposure rate in cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica, 45*(6), 694–703.

[毛秀珍，辛涛. (2013). 认知诊断 CAT 中项目曝光控制方法的比较．*心理学报, 45*(6), 694–703.]

Peng, Y. F., Luo, Z. S., Yu, X. F., Gao, C. L., & Li, Y. J. (2016). The optimization of test design in cognitive diagnostic assessment. *Acta Psychologica Sinica, 48*(12), 1600–1611.

[彭亚风，罗照盛，喻晓锋，高椿雷，李喻骏. (2016). 认知诊断评价中测验结构的优化设计．*心理学报, 48*(12), 1600–1611.]

Qi, S. Q., Dai, H. Q., & Ding, S. L. (2002). *Principles of modern educational and psychological measurement.* Beijing: Higher Education Press.

[漆书青，戴海琦，丁树良．(2002)．*现代教育与心理测量学原理*．北京：高等教育出版社.]

Reckase, M. D. (2003). *Item pool design for computerized adaptive tests*. Paper presented at the National Council on Measurement in Education, Chicago, IL.

Reckase, M. D. (2007). *The design of p-optimal item pools for computerized adaptive tests*. Paper presented at the 2007 GMAC Conference on Computerized Adaptive Testing.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127–141.

Tang, X. J., Ding, S. L., & Yu, Z. H. (2013). Application of computerized adaptive testing in cognitive diagnosis. *Advances in Psychological Science, 20*(4), 616–626.

[唐小娟，丁树良，俞宗火．(2012)．计算机化自适应测验在认知诊断中的应用．*心理科学进展, 20*(4), 616–626.]

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–361). Hillsdale: Lawrence Erlbaum Associates.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 51*(3), 337–350.

Tu, D. B., Cai, Y., & Dai, H. Q. (2013). A new method of Q-matrix validation based on DINA model. *Acta Psychologica Sinica, 44*(4), 558–568.

[涂冬波，蔡艳，戴海琦．(2013)．基于 DINA 模型的 Q 矩阵

修正方法. *心理学报, 44*(4), 558–568.]

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement, 48*(3), 255–273.

Xu, X. L., Chang, H. H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the the Annual Meeting of American Educational Research Association, Chicago, IL.

Yu, X. F., Luo, Z. S., Gao, C. L., Li, Y. J., Wang, R., & Wang, Y. T. (2015). An item attribute specification method based on the likelihood D^2 statistic. *Acta Psychologica Sinica, 47*(3), 417–426.

[喻晓锋, 罗照盛, 高椿雷, 李喻骏, 王睿, 王钰彤. (2015). 使用似然比 D^2 统计量的题目属性定义方法. *心理学报, 47*(3), 417–426.]

Zheng, C. J., & Chang, H. H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 40*(8), 608–624.

Optimization of test design for examinees with different cognitive structures

PENG Yafeng[1]; LUO Zhaosheng[1]; LI Yujun[2]; GAO Chunlei[1]

([1] *School of psychology, Jiangxi Normal University, Nanchang* 330022, *China*) ([2] *Center for Studies of Psychological Application/School of Psychology, South China Normal University, Guangzhou* 510631, *China*)

Abstract

Doctors have to use different medical technologies to diagnose different kinds of illness effectively. Similarly, teachers have to use well designed tests to provide an accurate evaluation of students with different cognitive structures. To provide such an evaluation, we recommend to adopt the Cognitive Diagnostic Assessment (CDA). CDA could measure specific cognitive structures and processing skills of students so as to provide information about their cognitive strengths and weaknesses.

In general, the typical design procedure of a CDA test is as follow: firstly, identify the target attributes and their hierarchical relationships; secondly, design a Q matrix (which characterizes the design of test construct and content); finally, construct test items. Within that designing framework, two forms of test are available: the traditional test and the computerized adaptive test (CAT). The former is a kind of test that has a fixed-structure for all participants with different cognitive structures, the latter is tailored to each participant's cognitive structure. Researchers have not, however, considered the specific test design for different cognitive structures when using these two test forms. As a result, the traditional test requires more items to gain a precise evaluation of a group of participants with mixed cognitive structures, and a cognitive diagnosis computer adaptive test (CD-CAT) has low efficiency of the item bank usage due to the problems in assembling a particular item bank. The key to overcome these hurdles is to explore the appropriate design tailored for participants with different cognitive structures.

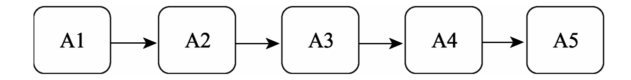
As discussed above, a reasonable diagnosis test should be specific for the cognitive structure of target examinees so to perform classification precisely and efficiently. This is in line with CAT. In CAT, an ideal item bank serves as a cornerstone in achieving this purpose. In this regard, Reckase (2003, 2007 & 2010) came up with an approach named *p*-optimality in designing an optimal item bank. Inspired by the *p*-optimality and working according to the characteristics of CDA, we proposed a method to design the test for different cognitive structures. We conducted a Monte Carlo simulation study to explore the different test design modes for different cognitive structures under six attribute hierarchical structures (Linear, Convergent, Divergent, Unstructured, Independent and Mixture).

The results show that: (1) the optimal test design modes for different cognitive structures are different under the same hierarchical structure in test length, initial exploration stage (Stage 0), accurately estimation stage (Stage 1); (2) the item bank for cognitive diagnosis computer adaptive test (CD-CAT) we built, according to the different cognitive structures' optimal test design modes, has a superior performance on item pool usage than other commonly used item banks no matter whether the fixed-length test or the variable-length test is used. We provide suggestions for item bank assembling basing on results from these experiments.

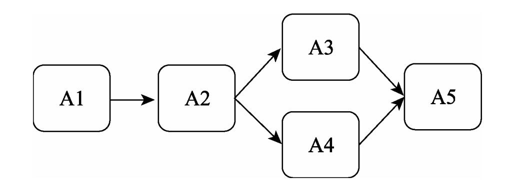
Key words cognitive structure; test design; item bank design

附录 1：六种基本的属性层级关系示意图(K=5)

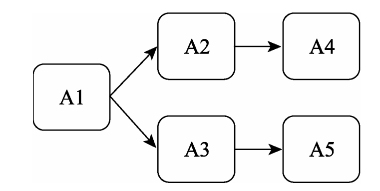
A. 直线型 (Linear)



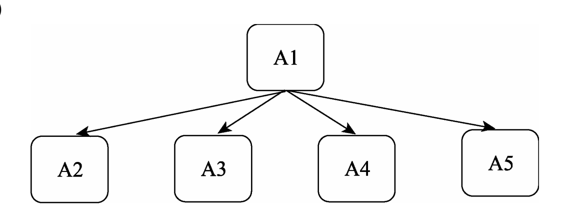
B. 收敛型 (Convergent)



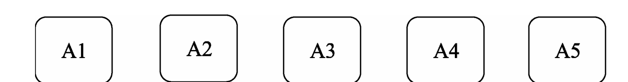
C. 发散型 (Divergent)



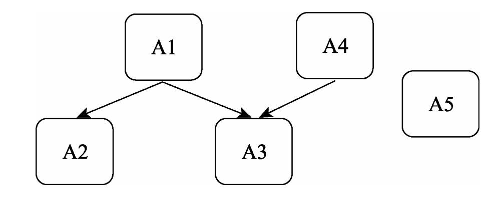
D. 无结构型 (Unstructured)



E. 独立型 (Independent)

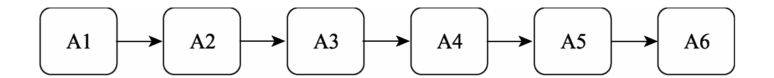


F. 混合型 (Mixture)

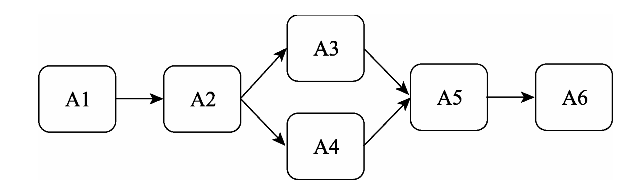


附录 2：六种基本的属性层级关系示意图(K=6)

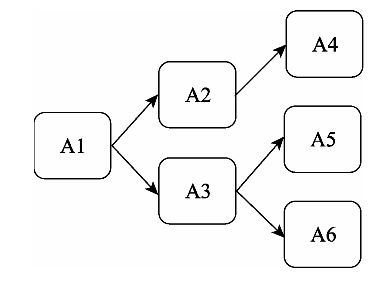
A. 直线型 (Linear)



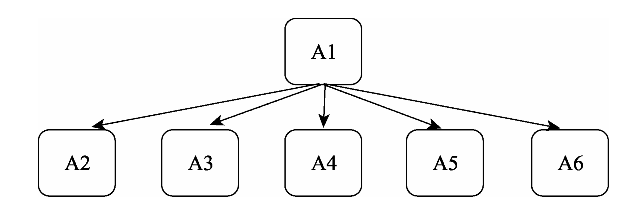
B. 收敛型 (Convergent)



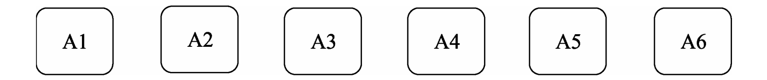
C. 发散型 (Divergent)



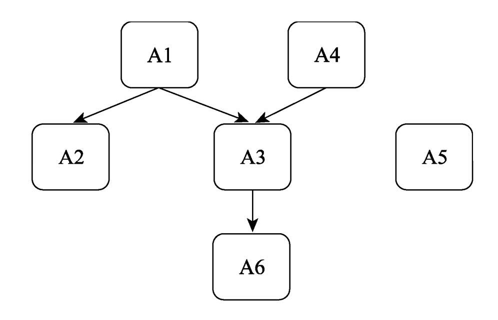
D. 无结构型 (Unstructured)



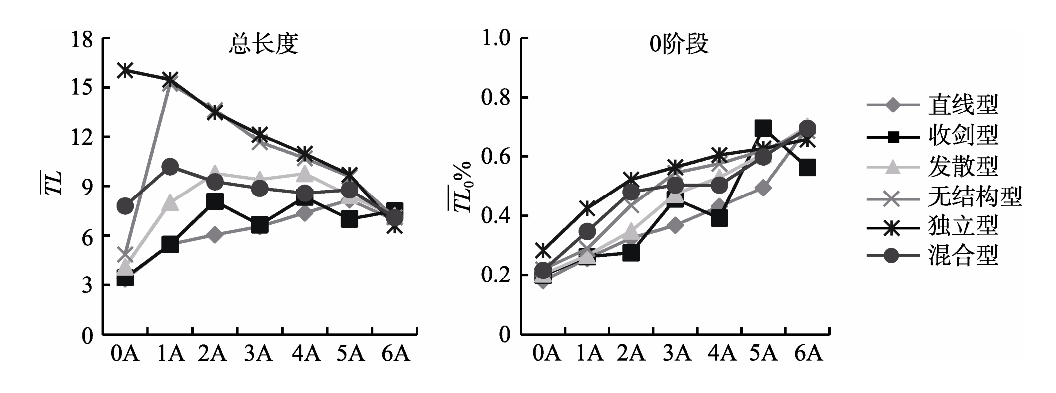
E. 独立型 (Independent)



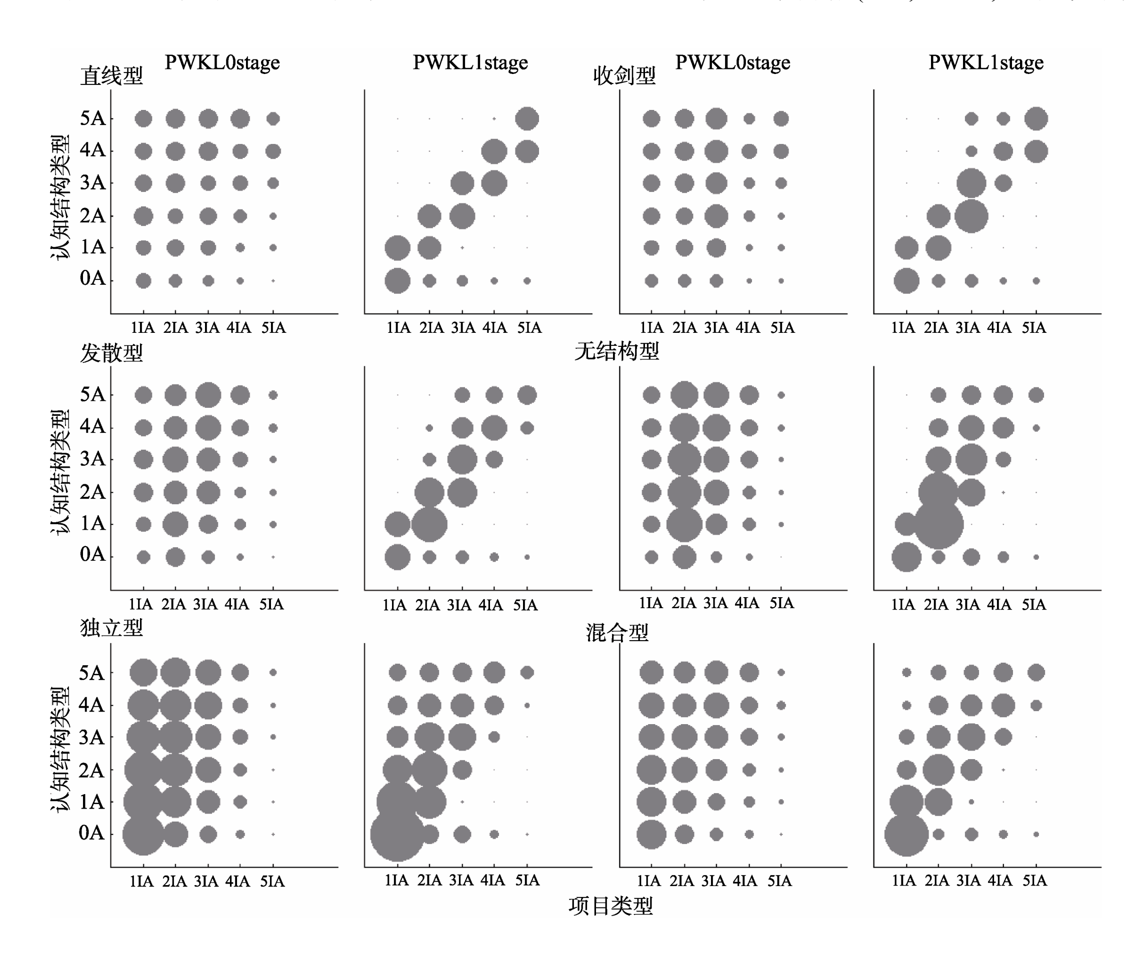
F. 混合型 (Mixture)



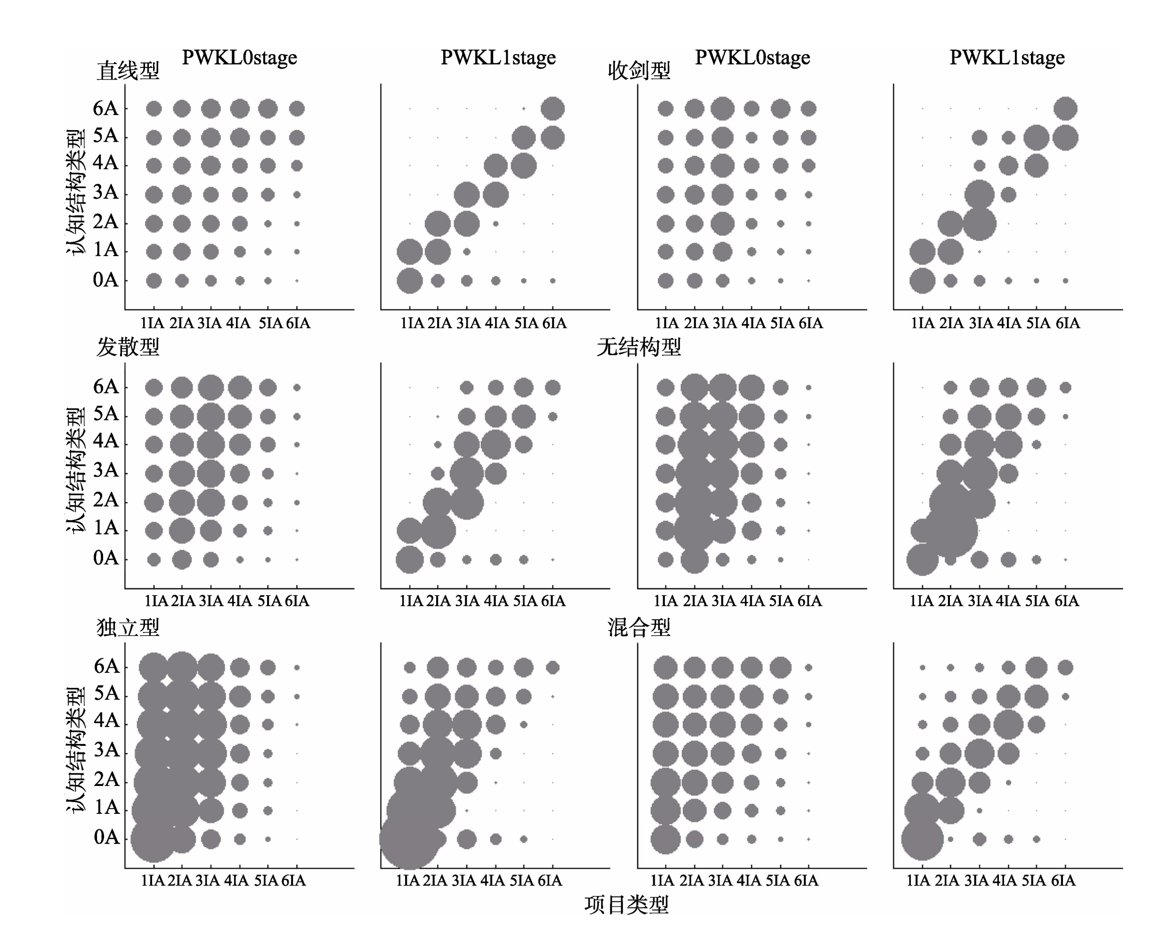
附录 3：六种属性层级关系下不同认知结构类型的平均测验总长度以及 0、1 阶段所占总长度百分比(K=6, 30 次平均结果)



附录 4：六种属性层级关系下所有认知结构在 0、1 阶段下选出的项目类型及其个数(K=5, $M+SD$, 30 次平均结果)



附录 5：六种属性层级关系下所有认知结构在 0、1 阶段下选出的项目类型及其个数(K=6, p_{90}, 30 次平均结果)



附录 6：六种属性层级关系下所有认知结构在 0、1 阶段下选出的项目类型及其个数(*K*=6, *M*+*SD*, 30 次平均结果)

